1.

International Journal of Energy and Statistics

Stochastic modelling techniques for generating synthetic energy demand profiles

Sandhya Patidar, David P. Jenkins and Sophie A. Simpson
Energy, Geoscience, Infrastructure and Society (EGIS); Heriot-Watt University; Edinburgh, UK

Received (received date)
Revised (revised date)
Accepted (accepted date)

This paper investigates three stochastic modelling procedures for generating N(user specified) synthetic annual electricity demand profiles at one-minute resolution. The paper reviews previous work in the application of Hidden-Markov modelling (HMM) for synthesizing highly stochastic time-series of domestic electricity demand through a sophisticated framework coalescing 480 distinct HMM. The efficiency of a proposed approach for integrating a time-series deseasonalizing technique with a single HMM has been studied in parallel with a compatible stochastic modeling framework of a time-series deseasonalized ARIMA model. Various statistical measures/characteristics of the real and synthetic profiles have been compared for all the three stochastic modelling procedures to identify the most efficient and practically suitable medium for generating synthetic electricity time-series at a fine temporal resolution. Results have been shown for both the individual buildings and the composite (aggregated) profiles of many buildings.

Keywords: Synthetic time-series; Energy; Hidden Markov Model; ARIMA Model; Time-series Deseasonalization.

Nomenclature

HMM	Hidden Markov Modelling.
$ARIMA$	Autoregressive Integrated Moving Average.
MA	Moving Average.
AR	Autoregressive.
$ARMA$	Autoregressive Moving Average.
EDA	Exploratory Data Analysis.
ACF	Auto-correlation Function.
$ARIES$	Adaptation and Resilience In Energy Systems.
$ARCC$	Adaptation and Resilience in a Changing Climate.

1. Introduction

Energy is often listed amongst priority issues in government policy [1]. As a result, a considerable amount of recent research investment has been made to promote quality research in the areas of energy. Researchers from different backgrounds are coming together to provide better understanding and solutions on various aspects of

energy use and efficiency. It has been anticipated that certain challenges in the area of energy are likely to develop further in the future [2] and substantial amount of research is required to develop efficient and sustainable energy related technologies and solutions to deal with this multi-faceted problem.

Appropriate data is essential to accurately record the extent of such a challenge, and researchers often face barriers in obtaining such data, relating to availability, costs, and resolution of datasets. Specifically, understanding the characteristics of energy demand that emanate from the built environment can require annual energy profiles at fine temporal resolutions (such as minute) for a large number of buildings. For example, a researcher might be interested in examining the effect a new appliance technology has on the electrical demand of an individual dwelling, but then also studying the impact of changes on the aggregated [a] electricity profiles, should that technology reach the mass market. For many cases, information required at such a large numerical scale (e.g. number of dwellings) and at such a fine resolution is not always available. To overcome such issues, statistical modelling techniques for generating synthetic profiles can be utilised.

Under this theme, in [3] the authors have previously presented novel application of a HMM technique in generating annual synthetic domestic electricity demand profiles at one-minute resolution through a complex framework of 480 integrated distinct HMM. The procedure presented in this reference was complex and had some identified limitations, particularly in capturing some stochastic characteristics of individual synthetic demand profiles (e.g. the visual inspection of synthetic profiles appeared to be more stochastic than the real profile, though other statistical characteristics were in good agreement). This related study presents and compares three methodologies to generate synthetic electricity demand profiles at one-minute resolution by exploiting stochastic modelling techniques, namely:

> *Modelling Procedure 1*: a framework of 480 integrated distinct HMM, presented in [3];
> *Modelling Procedure 2*: ARIMA modelling technique integrated with a time-series deseasonalization procedure; and
> *Modelling Procedure 3*: HMM modelling technique integrated with a time-series deseasonalization procedure.

The novelty of the present paper is the application of the time-series deseasonalization in association with a single HMM framework to design a systematic methodology for generating synthetic profiles, which can significantly simplify the methodology presented elsewhere [4], with the latter requiring the complex framework of 480 distinct HMM models. The proposed simplified procedure involving time-series deseasonalization has also been applied to a traditional ARIMA mod-

[a]Aggregated Profiles for a geographic area are usually constructed by integrating a large number of the individual energy demand profiles corresponding to many different buildings within the area.

elling framework. This paper will investigate the efficiency of these three stochastic modelling procedures.

Next section is intended to present a brief overview of the state of the art, discussing various stochastic modelling techniques that can be utilized for synthesizing electricity demand profiles. Section 3 presents an extensive EDA of the available data and thus formulates basis for the selection and designing of appropriate modelling techniques and framework. Details of the modelling procedure have been structured and presented in the Section 4 which is then followed by the Section 5 that provides a thorough analysis of model outcomes. Model performance has been systematically validated across individual profiles (Subsection 5.1), percentile distribution (Subsection 5.2) and ACF (Subsection 5.3). Practical application of the proposed methodologies has been discussed in the Section 6 which is then followed by the final discussion and conclusions in Section 7.

2. Background

Several techniques have been proposed in the past to generate synthetic electricity demand profiles. In context of the present paper, an extensive literature review for generating domestic electricity demand profiles can be found elsewhere [3, 5]. Some of the earlier modelling techniques for synthesizing electricity demand have limited applicability, primarily because most of these techniques require detailed information on various parameters which are not easily accessible in practice, such as activity profiles of the occupants, usage details of the various appliances, and lifestyle/habits of the occupants (see, e.g., [6–8]); secondly none of these techniques are known to be effective for synthesizing electricity demand at a temporal resolution of one-minute.

Domestic electricity demand profiles can be considered as a complex combination of highly stochastic processes and deterministic processes. Deterministic processes can be attributed to the steady operation patterns of certain appliances such as refrigerator, freezers, and televisions (and other appliances) on standby. However, there exist several factors such as usage pattern of appliances (depending on the number of occupants and their activities/habits), climatic influences (daylight hours can effect usage of lighting, loads on heating or cooling systems), which could be partly random and statistically suggests a more stochastic process.

A range of stochastic time-series modelling techniques such as moving average (MA), autoregressive (AR) [9], autoregressive-moving average (ARMA) [10], autoregressive integrated moving average models (ARIMA) (see, e.g., [11,12]), Monte-Carlo Simulation, Methods of Surrogate [13], Fast Fourier Transform [14], wavelet-based models [15], neural networks [16], and Markov Chain Models [17] have been explored to generate synthetic time-series across a wide range of discipline and interest areas. Moving average (MA), autoregressive (AR), and autoregressive-moving average (ARMA) modelling techniques are suitable for univariate stationary processes. Autoregressive integrated moving average models (ARIMA) are often used

for modelling stochastic non-stationary series. Through differencing, ARIMA models facilitate transformation of a non stationary time series in to a stationary time series and have numerous applications [18]. Some of the interesting applications of ARIMA modelling can be referred in time-series models involving seasonal effects, econometrics, and forecasting models.

HMM was first introduced by the [19] and, since then, has been successfully applied to a range of real world complex problems such as speech recognition [20], exploring time-sequential images [21], complex brain R images [22], prediction in computational biology [23] and modelling of protein structure [24] through large scale genomes-sequence analysis [25].

Statistically, ARIMA models exploit the correlation between an observation at time t with the subsequent observations in the past, i.e. *t-1, t-2, t-3, . . . , t-n*, whereas a HMM accounts for the evolution probabilities of the observed state at time t from one previous state (first order Markov model) at *t-1*, through a set of un-observed (hidden) states. This paper explored the detailed structural make-up of both the ARIMA model and the HMM and describe how these models can be efficiently framed for generating synthetic electricity profiles.

3. EDA

For a preliminary selection of an appropriate modeling technique, and to design a successful and efficient statistical modeling procedure, a key step is carrying out a detailed EDA of the original dataset. An EDA procedure reveals the dynamics and the detailed information on the statistical physiognomies of the dataset. Data available to conduct the proposed work consists of annual electricity demand profiles at one-minute resolution for nine different dwellings. Detailed information on the dataset can be referred elsewhere [26]. An in-depth analysis of the available dataset is presented and discussed below.

Table 1 presents the distribution of the average electricity consumption spread across the nine different dwellings over the year, spring (March-April-May), summer (June-July-August), autumn (September-October-November), and winter (December-January-February) period. It is interesting to note that the average annual consumption value along with the seasonal average consumption values varies considerably across the nine buildings (buildings are arranged in descending order for the average annual electricity consumption values of the dwellings). For all nine dwellings, the average electricity consumption during the summer and the spring period is moderately lower than the corresponding average annual consumption, with the summer consumptions slightly lower than the corresponding spring consumptions. Analogously, the average consumption during the autumn and the winter is moderately higher than the corresponding annual average consumption, with winter consumptions slightly higher than the corresponding autumn consumptions.

Table 2 presents a full spectrum of the five summary statistics (minimum, 25^{th} percentile, 50^{th} percentile, 75^{th} percentile, maximum) and two high-end extreme

Table 1: Average electrical consumption analysis (in kWh) for nine different real dwellings measured over the year, spring (March-April-May), summer (June-July-August), autumn (September-October-November), and winter (December-January-February) periods.

Building ID	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Annual	22.97	22.17	21.74	18.97	16.98	15.96	15.06	11.82	10.54
Spring	21.95	20.57	19.68	17.90	14.91	13.75	13.92	12.74	9.85
Summer	20.34	18.74	20.70	17.90	16.05	13.73	11.68	10.60	8.63
Autumn	24.60	22.18	21.44	19.81	17.01	18.04	16.50	11.01	11.44
Winter	25.31	26.54	25.27	20.23	20.08	17.89	17.91	12.66	12.00

percentile (90^{th} percentile and 95^{th} percentile) values measured across the nine different dwellings. These various statistics of annual consumption values for the nine different dwellings demonstrate considerable amount of variation. The table has been color-coded on a gray scale to expose any patterns in the percentile distribution of the electricity consumption values. It can be noted that for the percentile values lower than the 75^{th} percentile a consistent pattern in the electricity consumption values can be observed, with values decreasing from dwelling A to I. However, for the two high-end extreme percentiles (90^{th} and the 95^{th} percentile) a different pattern can be seen. It is interesting to note that the difference in the values observed at the 95^{th} percentile and the maximum is considerable. Specifically, these differences are 4.9 kW, 6.5 kW, 6.4 kW, 4.5 kW, 6.2 kW, 7.4 kW, 7.2 kW, 4.5 kW, and 6.3 kW for dwellings A, B, C, D, E, F, G, H and I respectively. This demonstrates, for such a high resolution dataset, that peak values can occur at very low frequency, being a result of a superposition of different activities/technologies that might only occur very rarely and for very short (e.g. a few minutes) periods of time.

Thus, the present analysis of the data suggests that up to 75% of the values in the data set are less than 1 kW and could be attributed to the standby loads, 20% of the values range between 0.8 to 3.2 kW and could be attributed towards a range of different appliance usages of the individual households used in a systematically predictable way, and the remaining 5% of the high-end values could be more difficult to predict events occurring in the households that do not follow a systematic trend (when, for example, comparing one dwelling to another dwelling). This demonstrates that key features of a domestic electrical demand profile (such as in Figure 2) relate to high values occurring for short time periods. Although these features are infrequent, relative to a typical day, they become important if the related activity (e.g. cooking appliances) correspond to an energy use that is likely to occur at the same time for many dwellings. This will result in a superposition of values that will cause features to be observed in an aggregated profile (discussed later).

Table 2: Five summary statistics (minimum, 25^{th} percentile, 50^{th} percentile, 75^{th} percentile and maximum) together with 90^{th} and 95^{th} percentile across a year of electricity demand data for nine different real dwellings.

Building ID	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Minimum	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25^{th} **Percentile**	0.6	0.5	0.4	0.5	0.3	0.3	0.3	0.2	0.2
50^{th} **Percentile**	0.7	0.6	0.6	0.7	0.4	0.5	0.4	0.4	0.3
75^{th} **Percentile**	1.0	1.0	0.9	0.9	0.6	0.7	0.7	0.6	0.5
90^{th} **Percentile**	1.6	1.6	1.8	1.3	1.4	1.0	1.3	0.9	0.8
95^{th} **Percentile**	2.9	2.9	3.0	1.7	3.2	1.8	1.7	1.2	0.9
Maximum	7.8	9.4	9.4	6.2	9.4	9.2	8.9	5.7	7.2
Color code	0.0-0.2		0.3-0.5		0.6-0.8		0.9-1.1		1.2 and over

The electrical load profile of a dwelling displays varied patterns across different days of a week, and months/seasons of the year, which can be partly correlated to the household type, occupant activity, climatic effects and other factors. To explore the dynamics of electricity load distribution corresponding to more random actions happening in the household, four percentiles (minimum, 75^{th} percentile, 90^{th} percentile and maximum) of minutely electricity load values are plotted and presented in Figure for three different dwellings (A, E and I) and for four different seasons (Spring -'March-April-Mayz', Summer -'June-July-August', Autumn -'September-October-November'and Winter -'November-December-January').

Figure 1 shows that the percentile distribution of the electricity load profile has considerable amount of variation across the three different dwellings, four seasons and time of the day. Visual inspection of the Figure 1 suggests that the percentile distribution of electricity consumption values between 00:00 - 08:00 hrs is slightly higher during the summer in comparison to the corresponding values for winter. This could be due to the warmer climatic conditions in summer putting extra loads on certain appliances such as refrigerator and freezers. Analogously, between 18:00 - 24:00 hrs the percentile distribution of electricity consumption values is slightly less during summer in comparison to the corresponding winter consumption values, which could be due to longer sunlight hours during the summer. This observation reflects the impact of climatic conditions on the electricity consumption. Again, the difference between the 95^{th} percentile value and the maximum value at each minute is similarly large across all the three dwellings and across all the four seasons.

Figure 2 provides a snapshot of the randomly selected one-minutely daily electricity demand profile for three dwellings A, E and I over two different day types (active and less active)[b] and two seasons (summer and winter) of a year. This

[b]Explained elsewhere [4], an active or less active weekday or weekend has been identified by statistically measuring activity levels happing in between 12:00 - 16:00 (this specific time period

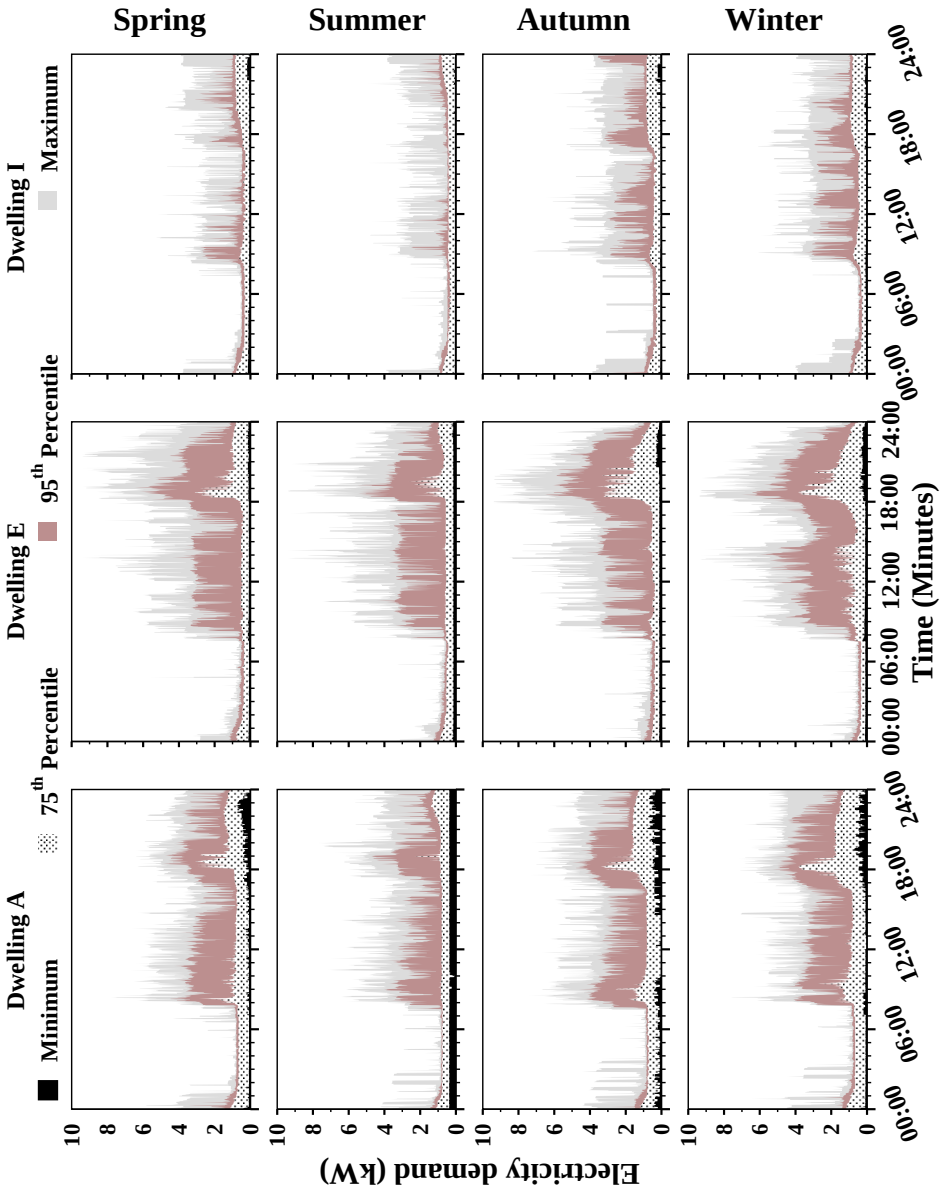

Fig. 1: Minimum, 75^{th} Percentile, 95^{th} Percentile and Maximum for three dwellings A, B and C, measured at one-minute resolution during spring (March-April-May), summer (June-July-August), autumn (September-October-November), and winter (December-January-February) period.

has been chosen to identify if occupants are in the house and active during the day time). A percentile analysis of the number of events (defined as energy demand exceeding 2kW) within the specified period suggested that a day could be labeled as 'active'if more than 10^{th} percentile values of such events happen within the specified period. Note: a different time period (than 12:00 - 16:00); a different value of threshold (than 2 kW); and a different level of percentile cut off (than 10^{th} Percentile) can be chosen to define diverse activity and occupancy patterns.

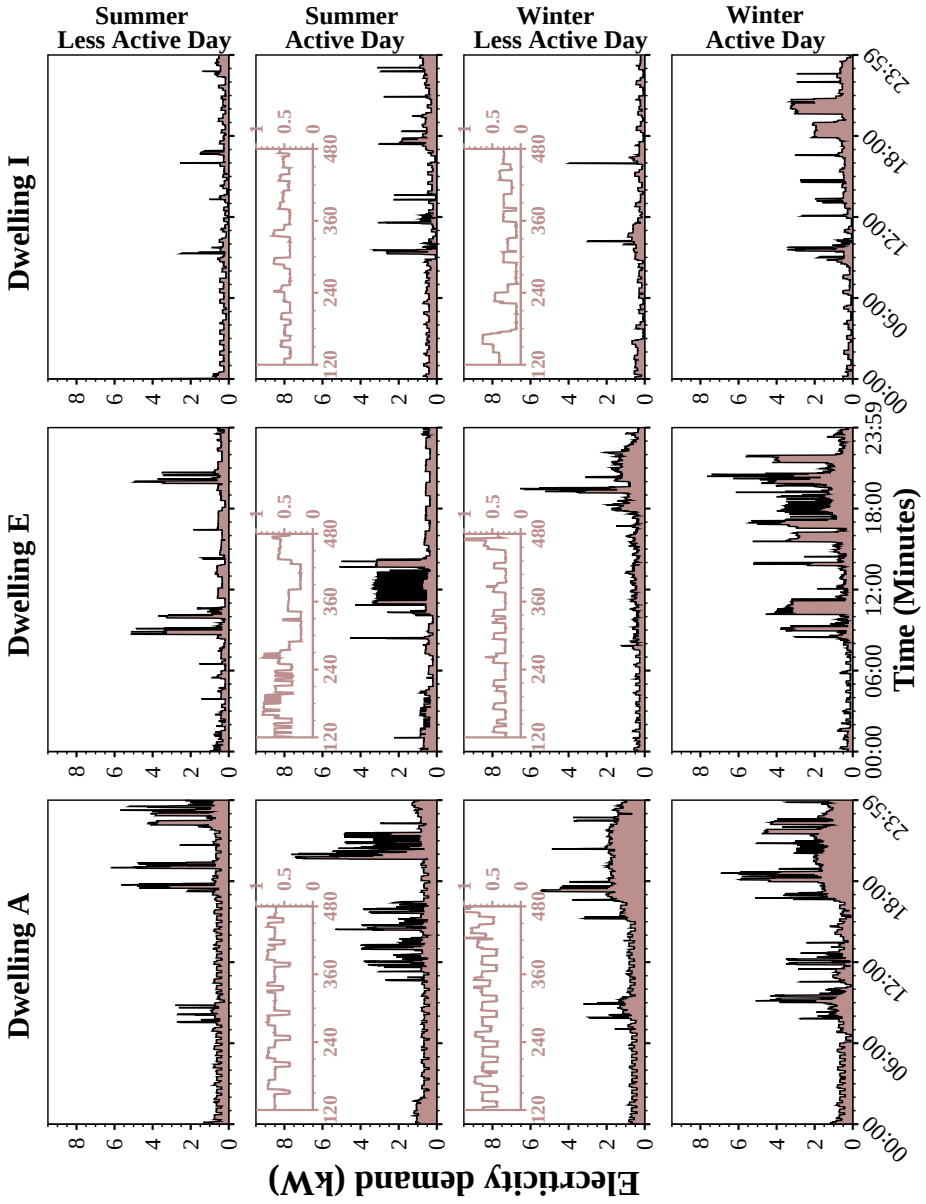

Fig. 2: Snapshot of randomly selected minutely real electricity demand profile for three dwellings A, E and I. One active and one less active day have been randomly selected during summer (June-July-August) and winter (November-December-January) period.

figure shows the key features of typical daily electrical load profiles and the variations in daily electricity use patterns across different dwellings, seasons, day types (active/less-active), and along different hours of the day. A typical daily electrical load profile is composed of randomly superimposed, 'standard base load cycles'[c] , 'time-specific loads'[d] and 'random loads'[e] . Thus, preliminary EDA concludes that the annual time series of the electricity consumption broadly demonstrates variation across: a) buildings; b) seasons; c) day type; and d) hours of a day.

4. Modelling Procedure

Synthetic time-series are designed to imitate the dynamics and the various statistical properties of the original time-series. This section presents step-by-step details of three distinct statistical modeling procedures to generate synthetic electricity demand profiles. All the three distinctive modeling procedures exploits the information collected through the EDA carried out in the above section to design an appropriate framework for replicating the dynamics and the key statistical features of the original series to the duplicate series.

4.1. *Modelling Procedure 1 - Integrated framework of 480 distinct HMM*

Markov models, in general, are suitable for problems involving temporal pattern recognition. They survey the original dataset to identify temporal patterns in the data and then simulate the patterns to extrapolate a desired number of synthetic series. Markov models are known to have limited success when applied to a complex dataset (see below for what such a dataset might entail). However, for complex datasets HMM technique appeared to provide better solution than the conventional Markov modeling technique. This section presenta the methodology designed to exploit the Hidden-Markov Model (HMM) through a well-organized framework.

HMM can be identified as a probabilistic modeling toolkit for modeling complex datasets, which can be categorized as a finite set of observed states with each observed state comprised of countable number of underlying observations represented as hidden states of the system. The structural composition of HMM consist of five basic components:

(1) **State space** – A finite set of well-defined observed states.

[c]Standard base load cycles - Mainly attributed to continuously running appliances such as refrigerator, freezers TV/Laptops on standby mode

[d]Time-specific loads - Depending on the occupancy and occupants' life style certain electrical appliances (such as lighting, dishwashers, washing machines, cooking, TV/Games, etc.) could have a consistent usage pattern during specific times of the day.

[e]Random loads: Quite often depending on household type, life-style and habits of occupants' climatic conditions, usage of electrical appliances could be partly random and attributed to the random load.

(2) **Hidden states space** – A finite set of underlying observations (often referred as 'hidden states') corresponding to each observed states.
(3) **State transitional probability matrix** – A matrix containing transitional probabilities between different observed states.
(4) **Emission probability matrix** – A matrix containing emission probabilities of underlying observations from the observed states.
(5) **Initial probability set** – A set defining initial probabilities of observed states.

Analysis of the key statistical features of the electricity demand data clearly indicates that time series of the electricity demand profiles embrace variability across the different dwellings, seasons/months, day type and hour of the day. Therefore, to initialize simulations of synthetic time-series, the original electricity demand dataset has been carefully segregated into 480 distinct parts and a distinct HMM has been fitted to each part. Segregation procedure of the data can be summarized into three stages, described below:

- **Stage 1** To account for effects of seasonal variability, the annual time series is segregated into six parts by pairing data from two consecutive months together i.e. January- February; March-April; ... ; November-December.
- **Stage 2** To capture variability arising due to activity patterns, each of these six parts is then divided into four day types, namely active weekday, active weekend, less-active weekday, and less-active weekend.
- **Stage 3** To incorporate variability at different periods of a day, each day has been divided into 20 distinct hour types. EDA suggests that the period from 00:00 to 08:00 can be divided into four parts (each of 2 hrs) and then remaining period from 08:00 to 24:00 can be divided into 16 distinct parts (each of 1 hr).

Thus, by segregating an annual time-series of minutely electricity demand profiles into $6 \times 4 \times 20 = 480$ parts and then by fitting the above specified HMM to each of these 480 parts through an efficient algorithm compiled in R (Statistical programming language and interface) [27] an integrated framework of 480 HMM has been designed. The HMM model is fitted in R by exploiting the HMM-package [28]. For annual electricity demand profiles, the above specified five basic components of HMM were defined as follow:

(1) **State space** – a set of five distinct observed states (A, B, C, D, E)[f] has been defined by conducting a systematic percentile analysis of all observed values of annual electricity demand profiles.

[f]Explained elsewhere [4], State A is defined as load between the minimum and 20^{th} Percentile value; state B as load between the 20^{th} and 40^{th} percentile value; state C as load between the 40^{th} and 60^{th} percentiles value; state D as load between the 60^{th} and 80^{th} percentile value; and state E as load between the 80^{th} percentile and maximum value.

(2) **Hidden states space** – a set of underlying (hidden) states corresponding to each observed state has been defined by accounting all discrete values with one decimal place lying within the range of observed states. Therefore if, for a dwelling, state A corresponds to the event of observing a load value between 0 and 1 kW then a set of underlying hidden states corresponding to observed state A will be $0.1, 0.2, \ldots, 0.9$ kW load values; similarly, if state B corresponds to the event of observing a load value between 1 and 2 kW then set of underlying hidden states will $1.1, 1.2, \ldots, 1.9$ kW load values; and ... so on.
(3) **State transitional probability matrix** containing transitional probabilities of five different observed states has been identified as a 5×5 matrix, say S, where S_{ij} (elements in i^{th} row and j^{th} column) giving probability of transition from state i to state j.
(4) **Emission probability matrix** containing emission probabilities of underlying (hidden) observations from the discrete observed states has been identified as a5×9 matrix, say E, where E_{ij} (elements in i^{th} row and j^{th} column) giving probability of transition from observed state i to underlying (hidden) state j.
(5) **Initial probability set** as a set of five initial probabilities corresponding to each observed state has been obtained from the analysis of the time series.

Modelling Procedure 1 can be simulated to generate any number of user-specified, say N (integer number), synthetic annual electrical demand profile at minutely resolution.

4.2. *Modelling Procedure 2 - ARIMA model with time-Series deseasonalization*

Generally, time-series can be decomposed into three components: trends, seasonal variation and random components. For effectively modelling a time series it is desired that any trends, seasonal variation, random components and residual correlations between the data should be appropriately accounted for in the modelling procedure. Annual one-minutely electricity demand profiles available for the present work are associated with seasonal variation transversely through the daily, weekly and monthly cycles. Therefore, it is important that the modelling procedure should effectively account for these seasonal variations in the synthetic time-series.

Autoregressive modelling is a commonly approached technique for generating synthetic univariate time series. $ARIMA(p, d, q)$ models are generally encouraged for modelling non-stationary processes, where p denotes delay, q denotes moving average and d denotes the order of differentiation required for transforming non-stationary series into stationary series [17]. A time-series deseasonalization technique promotes segregation of seasonal variations from the series while leaving behind any trends and random components in the residual data. A systematic pro-

cedure for deseasonalizing a time-series can be referred elsewhere [29]. To generate a synthetic time-series of annual electricity demand under modelling procedure 2, the time-series deseasonalization technique has been integrated with the ARIMA models. The modelling procedure is consists of two parts and is described as follows:

Step I to III involves deseasonalization of the annual electricity demand profiles (described as below):

Step1 Various methods are available for removing trend component from the time-series, one of the simple approach is taking logarithm . Thus, step 1 involves log-transforming time-series of annual electricity demands, say time-series

$$\{E_t\} = E_{t_1}, E_{t_2}, \ldots, E_{t_i}, \ldots, \tag{1}$$

to the log-transformed series

$$\{\log(E_t)\} = \log(E_{t_1}), \log(E_{t_2}), \ldots, \log(E_{t_i}), \ldots. \tag{2}$$

Step2 To remove seasonal effects at an hourly scale, hourly averages and hourly standard deviations of the log-transformed series are computed, i.e.

$$\{E\}_{\mu_{hr}} = E_{\mu_1}, E_{\mu_2}, \ldots, E_{\mu_i}, \ldots, \tag{3}$$

and

$$\{E\}_{\sigma_{hr}} = E_{\sigma_1}, E_{\sigma_2}, \ldots, E_{\sigma_i}, \ldots, \tag{4}$$

where

$$\begin{aligned} E_{\mu_1} &= \mathit{Average}\quad \mathit{of}\quad (\log(E_{t_1}), \log(E_{t_2}), \ldots, \log(E_{t_{60}}))\,, \\ E_{\mu_2} &= \mathit{Average}\quad \mathit{of}\quad (\log(E_{t_{61}}), \log(E_{t_{62}}), \ldots, \log(E_{t_{120}}))\,, \\ &\ldots \quad \mathit{so}\quad \mathit{on}. \end{aligned} \tag{5}$$

$$\begin{aligned} E_{\sigma_1} &= \mathit{standard}\quad \mathit{deviation}\quad \mathit{of}\quad (\log(E_{t_1}), \log(E_{t_2}), \ldots, \log(E_{t_{60}}))\,, \\ E_{\sigma_2} &= \mathit{standard}\quad \mathit{deviation}\quad \mathit{of}\quad (\log(E_{t_{61}}), \log(E_{t_{62}}), \ldots, \log(E_{t_{120}}))\,, \\ &\ldots \quad \mathit{so}\quad \mathit{on}. \end{aligned} \tag{6}$$

Step 3 To deseasonalize the time series, the hourly averages are subtracted from each minutely values of the log-transformed series and then divided by the difference of hourly standard deviation values, i.e.

$$\{E\}_{Deseasonalize} = E_{Deseasonalize}(t_1), E_{Deseasonalize}(t_2), \ldots, E_{Deseasonalize}(t_i), \ldots, \tag{7}$$

where

$$
\begin{aligned}
E_{Deseasonalize}(t_1) =& \frac{\log(E_{t_1}) - E_{Avg_1}}{E_{sd_1}}, \ldots, E_{Deseasonalize}(t_{60}) = \frac{\log(E_{t_{60}}) - E_{Avg_1}}{E_{sd_1}}, \\
E_{Deseasonalize}(t_{61}) =& \frac{\log(E_{t_{61}}) - E_{Avg_2}}{E_{sd_2}}, \ldots, E_{Deseasonalize}(t_{120}) = \frac{\log(E_{t_{120}}) - E_{Avg_2}}{E_{sd_2}}, \\
& \ldots \quad so \quad on.
\end{aligned}
\tag{8}
$$

Step 4 involves fitting of the $ARIMA(p, d, q)$ model to the deseasonalized time series of annual electricity demand. The ARIMA model has been fitted in 'R'by exploiting Forecast-package [30]. Simulation of the ARIMA model n times generates n synthetic deseasonalized annual electricity demand time series of minutely resolution. To introduce seasonal variation back into the series, the reverse procedure of the above step has been followed, i.e. these synthetic deseasonalized time series were multiplied by the hourly standard deviation and then hourly averages were added accordingly. Finally, the exponential of the series terms were taken which provides the required n synthetic annual electricity demand profiles at one-minute resolution.

4.3. *Modelling Procedure 3 - HMM model with time-Series deseasonalization*

Modelling Procedure 3 combines a HMM with the above specified time-series deseasonalization procedure. To the deseasonalized series obtained as described above a HMM model is fitted in R using HMM-package [28] and then exactly the same procedure as described in Part II has been followed to add seasonal trends and construct the required n synthetic annual electricity demand profiles at one-minute resolution.

It should be noted that Modelling Procedure 1, which proposes fitting of 480 distinct HMM to the different partitions of the original time-series, is complex and involves intense data processing, whereas Modelling Procedure 3 requires a single HMM model. At this point, it can be argued that Modelling Procedure 2 and 3 are comparatively much simpler, cost effective (in terms of computational resources and time required) and easier to implement in practical applications than Modelling Procedure 1, though the fact remains that the success of a model depends on its efficiency in generating a desired outcome of high quality. The efficiency of Modelling Procedure 1 in generating synthetic profiles has been rigorously tested and validated in [3]. This paper thoroughly investigates the efficiency and limitations of Modelling procedure 2 and 3 in comparision to the Modelling procedure 1.

5. Results

This section examines the results of the synthetic electricity demand profiles constructed from the three modelling procedures (1, 2 and 3). To serve this purpose, various statistical characteristics of the synthetic electricity demand profiles were compared with the real electricity demand profiles.

5.1. *Individual profiles*

In section EDA, Figure 2 displayed the variation in usage patterns and key dynamical characteristics of a typical daily electricity load profile for three dwellings A, E and I, across two different day types (active and less active) and two seasons (Summer and Winter). Figure 3 provide a quick visual examination to assess the success of modeling procedures 1, 2 and 3 (respectively) in simulating these variations for dwelling A [please refer to Appendix Figure 8 and Figure 9for dwelling E and I respectively].

As already discussed, a typical daily electricity load profile can be considered as being composed of 'standard base load cycles', 'time specific loads'and 'random loads'. Although quite specific features will occur due to dwelling type, occupant lifestyle and climatic conditions, some basic features of different profiles can still be compared.

A quick inspection of Figures 3, 8, and 9 reveals that all the three modelling procedures are reasonably successful in capturing most of the basic features of the one-minutely electricity load distribution and observed variations of the electricity load profiles across the different hours of the day, day types, season types and the dwelling types. However, a closer inspection reveals that the general makeup of these synthetic profiles is slightly different across the three modelling procedures. To demonstrate this difference, the inset plots of Figures 2, 3, 8, and 9, highlighting the standard base load cycles, have been used. Standard base load cycles constructed through modelling procedure 1 have comparatively more irregular patterns of peaks and troughs than modelling procedures 2 and 3. Modelling procedure 2 generates more frequently appearing sharp peaks than modelling procedures 1 and 3, which is less representative of actual cycles in real data (which would be derived from fridge/freezers). Similarly, on comparing peak loads (exceeding 2 kW values), synthetic profiles generated by modelling procedure 1 display considerably more irregularity of peak load; whereas modelling procedure 2 generates load profiles with comparatively sharper and more frequent peak load values. The overall makeup of one-minute electricity load profiles generated by modelling procedure 3 seems to be slightly more consistent with the real load profiles.

5.2. *Percentile distribution*

Visual inspection of the minutely electricity load profiles provides evidences to support the effectiveness of the modelling procedures in replicating various statistical

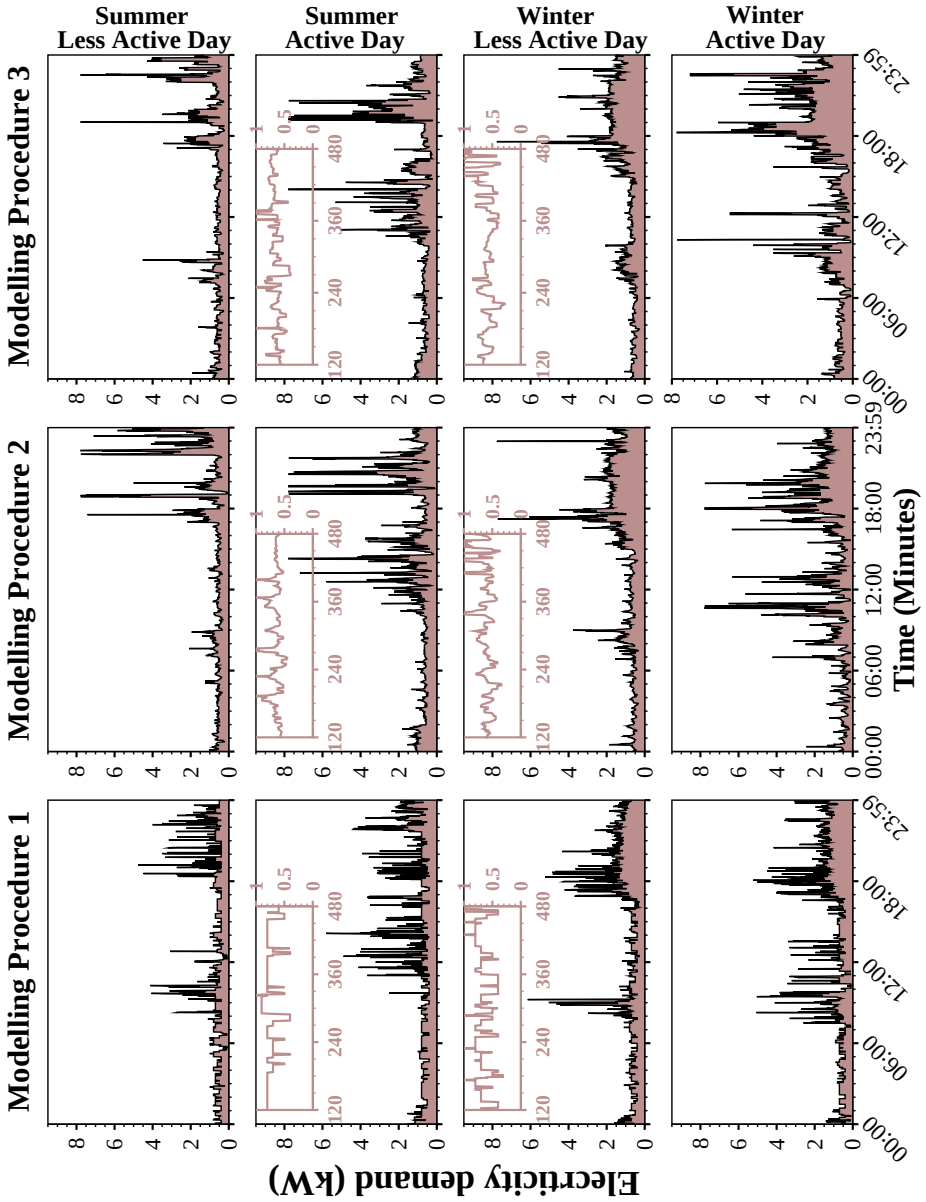

Fig. 3: Snapshot of randomly selected minutely synthetic electricity demand profile for dwelling A generated by three modelling procedures 1, 2 and 3. One active and one less active day have been randomly selected during summer (June-July-August) and winter (November-December-January) period.

characteristics of the real profiles. This section presents a more thorough comparison of the percentile distribution (10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, 95^{th} and Maximum) for the real dwellings A (Figure 4), [Please refer to Appendix Figure 10 and Figure 11 for dwelling E and I respectively] with the corresponding percentiles at one-hour resolution for randomly selected synthetic annual profiles generated by the modelling procedure 1, 2, and 3.

From the analysis of Figures 4, 10, and 11 all three modelling procedures are successful in capturing dynamics of all the percentiles lower than the 75^{th}, (though modelling procedure 1 appears to display slight variations for dwelling E and I). This effect is specifically noticed during the active period[g], i.e. between the morning (07:00 - 10:00) and the evening (17:00 - 24:00) hours. During these hours the percentile values of real profiles are considerably higher than those of the synthetic profiles generated by modelling procedure 1. Additionally, the maximum load values estimated by modelling procedure 1 is consistently lower than the value observed for the real profiles, whereas for modelling procedures 2 and 3 the maximum load values are slightly higher than the values observed for the real profiles. Overall it is suggested that modelling procedures 2 and 3 are performing better than modelling procedure 1; and modelling procedure 3 looks slightly more effective than modelling procedure 2 in capturing percentile distribution of the real profiles.

5.3. *ACF*

Autocorrelation, also referred as 'lagged correlation'or 'serial correlation', of a time series can be defined as the correlation of the time-series with its own past and future values. Autocorrelation can be considered as a signature property of time-series data, which detects if successive terms of the series are independent of each other. A series with no autocorrelation is a random series and therefore cannot be predicted whereas positive/negative autocorrelation of values indicates some certain degree of association between the observations. To detect autocorrelation in a time-series, i) time-series plots; ii) lagged scatter plot and iii) ACF plots (also known as 'correlogram'[h]) are commonly used.

Figure 5 displays ACF plots of the electricity demand profiles for dwellings A, E and I[*represented by the solid black line with the solid circles*] compared with the ACF plots of the randomly selected synthetic annual series generated by modelling procedure 1 [*solid gray (orange online) line with solid diamonds*], 2 [*solid gray (brown online) line with solid squares*], and 3 [*solid gray line with cross*]. From the analysis of Figure 5 it can be noticed that for all the three dwellings, the ACF for the synthetic series generated by modelling procedure 3 is bordering closest to the ACF of the real series, whereas the ACF of modelling procedure 1 is running to the

[g]Most of the peak load values (exceeding 2kW) are observed during this period of the day.
[h]Correlogram - plots correlation of the data (y-axis) for the various lags (x-axis). For example: at lag 1 ACF represents correlation of observation at time t and one previous period $t-1$. Similarly at lag h ACF represents correlation of observation at time t and at h previous periods $t-h$

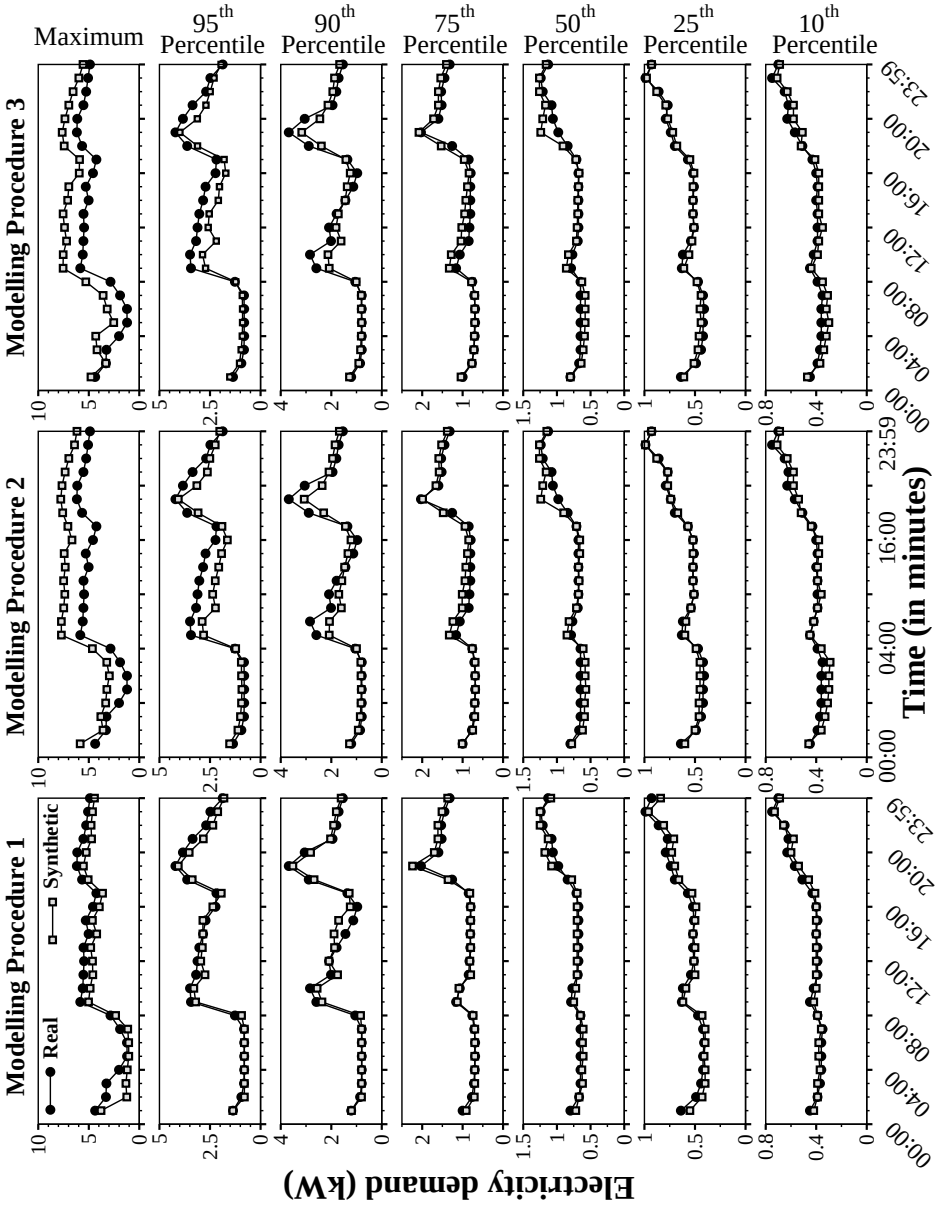

Fig. 4: Maximum and 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th} and 95^{th} percentile at one-hour resolution over the entire year period. Percentiles are compared for real dwelling A with percentiles measured for randomly selected synthetic annual profiles generated by modelling procedure 1, 2 and 3.

outermost of the real series. Thus analysis of ACF clearly indicates that modelling procedure 3 is the most effective procedure in capturing the correlation trends of the successive terms of the real series.

In summary, from the analysis presented in this section all the three modelling procedures appear to be compatible for generating synthetic electricity demand profiles at one-minute resolutions. However, when considering the fine details of the comparison analysis, modelling procedure 3 appears considerably more effective than modelling procedures 2 and 1 (with modelling procedure 1 the least effective).

6. Practical Application

Synthetic electricity demand profiles of the type described in this paper have a range of possible applications, including the construction of aggregated electricity load profiles, energy demand forecasting and compensating for missing data in empirical datasets. This section will demonstrate a significant application of the synthetic profiles in constructing aggregated profiles. Aggregated electricity demand profiles are particularly useful when considering the impact of technology and/or policy changes on how energy is supplied. The management of the electricity grid, for example, is mostly concerned with the demand profile of large regions of building stock (as well as nationally), rather than individual demand profiles of single dwellings.

Figures 6 and 7 demonstrate the efficiency of the proposed modelling procedures 1, 2, and 3 in capturing the transient variations of aggregated electricity load profiles. Figure 6 compares the aggregated profiles of the nine real dwellings [*solid black lines*] at one-minute resolution with the aggregated profiles of 225 synthetic dwellings [*solid brown line*]. The 225-dwelling plots were synthesized by creating 25 synthetic profiles from each real dwelling, and applying modelling procedures 1, 2, and 3 (in Figure 6 top row, middle row and bottom row respectively) for three different versions of this aggregated synthesis. These 225 aggregated synthetic profiles and the aggregated profiles of the nine dwellings are then compared to a randomly selected summer day of 13^{th} June (in Figure 6 left panel) and winter day 17^{th} Jan (in Figure 6 right panel). Key observations from this analysis are: i) The synthetic aggregated electricity profiles are smoother than the individual profiles - as would be seen with empirical data - and the increase in the number of the distinct profiles entering in the aggregation process increases the smoothness of the final aggregated profile (an effect often referred to as After Diversity Maximum Demand when observed in real electricity demand data). Thus, the aggregated profiles constructed from 225 synthetic profiles are much smoother than the 9 aggregated profiles corresponding to the real profile; ii) The aggregated profiles of 225 synthetic dwellings follows the same trends as of the aggregated real profiles of nine dwellings across the different hours of the day and also for different seasons (summer and winter).

Figure 7 compares real substation profiles of 230 dwellings [*solid black lines*] at ten-minute resolution with the aggregated profiles of 225 synthetic profiles [*solid*

brown line]. This substation data was available at 10-minute resolution and therefore minutely synthetic aggregated profiles are averaged together to get comparable data at the same resolution, used for the analysis illustrated in Figure 7. The left panel and right panel display analysis for the above randomly selected days of 13^{th} June and 17^{th} Jan respectively. Top row, middle row and bottom row respectively displays aggregated synthetic series constructed through the application of modelling procedure 1, 2, and 3.

It is important to note that the substation data is of a different group of buildings and different year than the data that is used to synthesis the individual dwellings comprising the 225-dwelling profile. There are therefore a number of reasons for the profiles not to exactly match. One such reason is that the more recent substation data is taken from homes with a higher penetration of low energy lighting; this would result in lower peak demand during the evening for the real substation data. This is evident in the three winter profiles of Figure 7. The comparison is therefore carried out to identify whether a similar scale of variation is being achieved by the synthesized profiles, but further work is required to validate this. The current validation does indicate that the difference between summer and winter profiles exhibits approximately similar patterns in the synthetic data as seen in the real data. Timing of peak demand is also predicted at similar times, which would be crucial to an energy supplier. This is further explored by the authors [3] in a more application-focused study .

7. Conclusion and discussion

Energy demand across the globe is rising rapidly and it is important to balance an efficient energy demand and supply chain. To conduct research into the various factors likely to influence future energy demand, such as impact of new technologies, climate change, and government policies, high quality data at an appropriate resolution is needed. However, in practice it could be challenging to gain access to such data for conducting these studies. This could limit our understanding of how different demand-side technologies, if applied across a large proportion of the building stock, might impact the supply of that energy in terms of magnitude and timing of peak demand and other demand profile characteristics.

This paper proposed three stochastic modelling strategies for generating synthetic electricity demand profiles. All three proposed techniques were critically analysed compared and validated. For generating synthetic annual energy demand profiles at one-minute resolution, modelling procedure 1 proposes a sophisticated framework of 480 integrated HMM. Modelling procedures 2 and 3 propose much simpler and efficient solutions by integrating time-series deseasonalization techniques with ARIMA and HMM approaches.

The validation included the analysis of key statistical features of the real and synthetic electricity demand profiles. It was found that all three techniques are able to generate synthetic electricity demand profiles at one-minute resolution. However,

modelling procedure 3 appears to provide a simpler and more efficient solution in comparison with the other two models.

Modelling procedure 2 uses an ARIMA model at its core whereas modelling procedures 1 and 3 adopt an HMM approach. Statistically, ARIMA models are based on exploiting the correlation property of the series with itself whereas the HMM technique exploits the probability of a system to jump from one state to another state. Therefore, ARIMA models appear to capture most of the statistical characteristic of real series but are not highly successful in emulating the basic structure of electricity load profiles (specifically, with the data presented here, for peak values of over 2 kW). Consequently, it has been noted that individual profiles generated through the application of modelling procedure 2 exhibit sharper peaks, which appear to be unrealistic when compared to real data. Although both modelling procedures 1 and 3 use HMM at their core, it should be noted that modelling procedure 3 is much simpler to implement in practice than modelling procedure 1. This would have clear computational advantages if adopted in industry.

Work is in progress to generalize the stochastic modelling techniques of generating synthetic profiles for wider application areas. In the Adaptation and Resilience In Energy Systems (ARIES) project [31], the possible generalization of the technique in synthesizing demand profiles for non-domestic buildings could also be investigated. Plans are in place to include larger datasets from real dwellings to examine the robustness of the technique and application of the methodology in compensating missing datasets. Various interesting approaches are available in literature for methodologically decomposing and forecasting a time series such as Empirical Mode Decomposition, Singular Spectrum Analysis, Exponential Smoothing etc., which can be investigated for expanding this work in future. Nevertheless, the present work presents a novel and applicable technique for effectively simulating highly stochastic time-series at a fine temporal resolution of one-minute.

Acknowledgements

The ARIES project is funded by the Engineering and Physical Sciences Research Council (Grant Ref. EP/I03534X/1) as part of the Adaptation and Resilience in a Changing Climate (ARCC) programme.

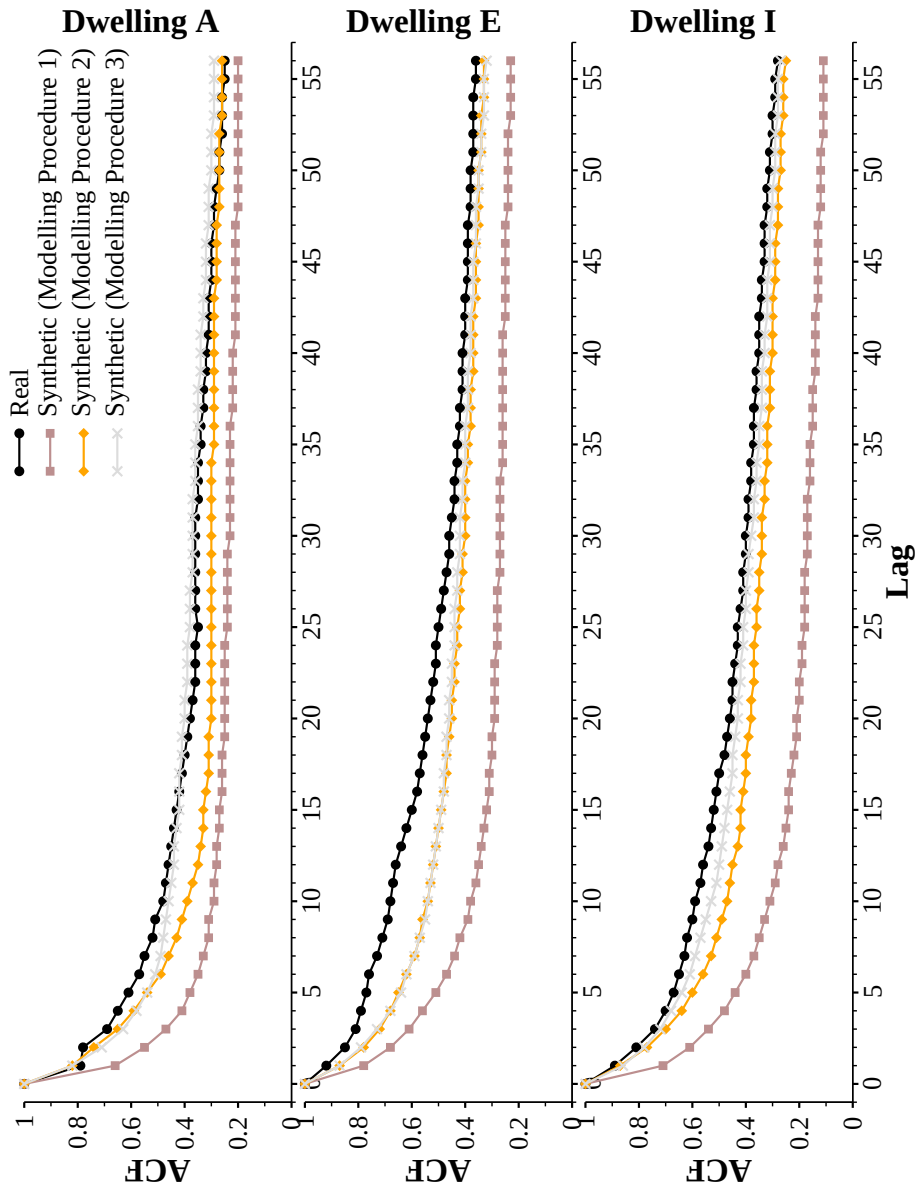

Fig. 5: Comparing ACF for real dwelling A, E and I[represented by solid black line with solid circles] with auto-correlation function measured for randomly selected synthetic annual profiles generated by 'modelling procedure 1'[*solid gray (orange online) line with solid diamonds*], 'modelling procedure 2'[*solid gray (brown online) line with solid squares*], 'modelling procedure 3'[*solid gray line with cross*].

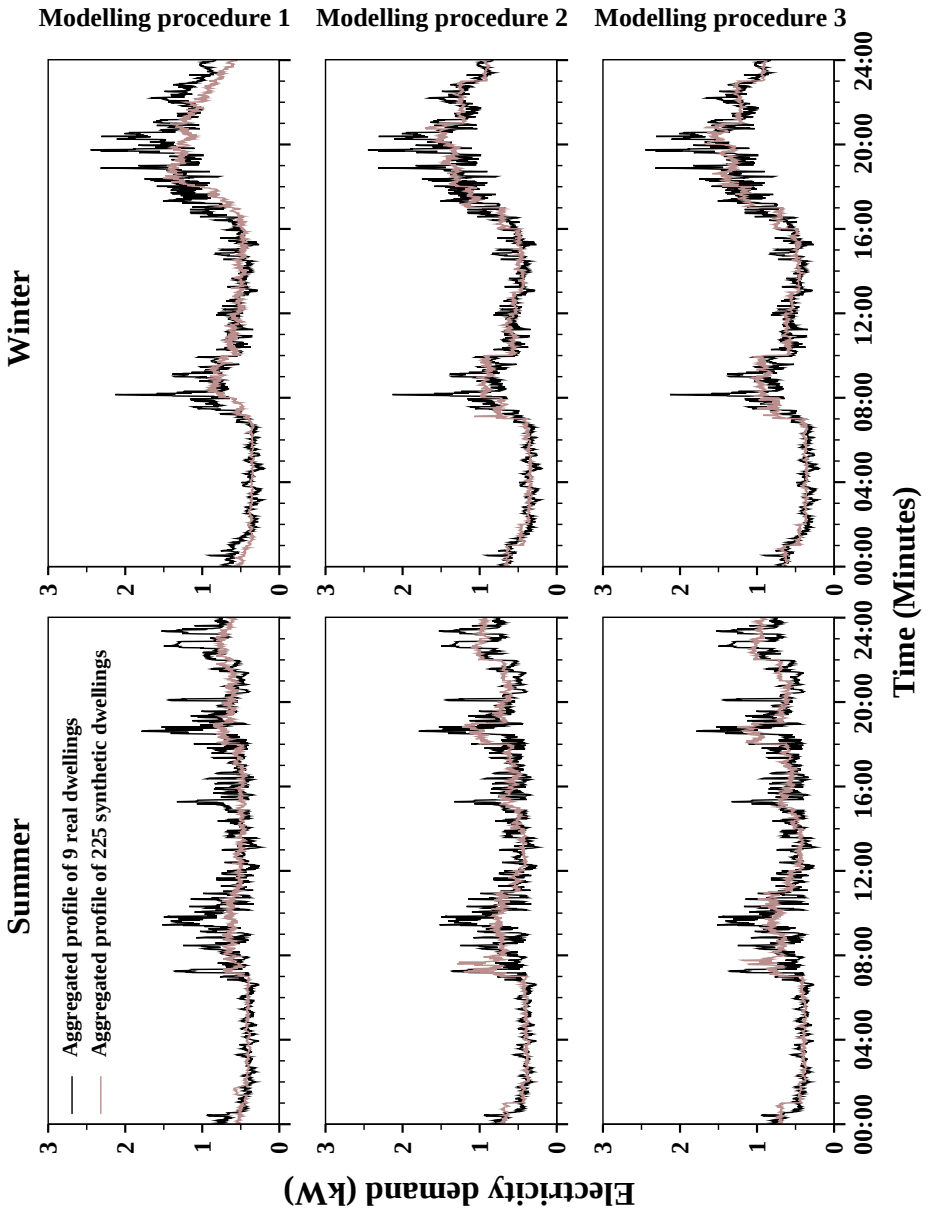

Fig. 6: Aggregated profile of nine dwellings [*solid black lines*] at one-minute resolution compared with aggregated profile of 225 synthetic dwellings [*represented by solid brown line*] for a randomly selected summer day 13^{th} June (**Left Panel**) and for a randomly selected winter day 17^{th} Jan (**Right Panel**). Top row 'modelling procedure 1'; middle row 'modelling procedure 2'; bottom row 'modelling procedure 3'.

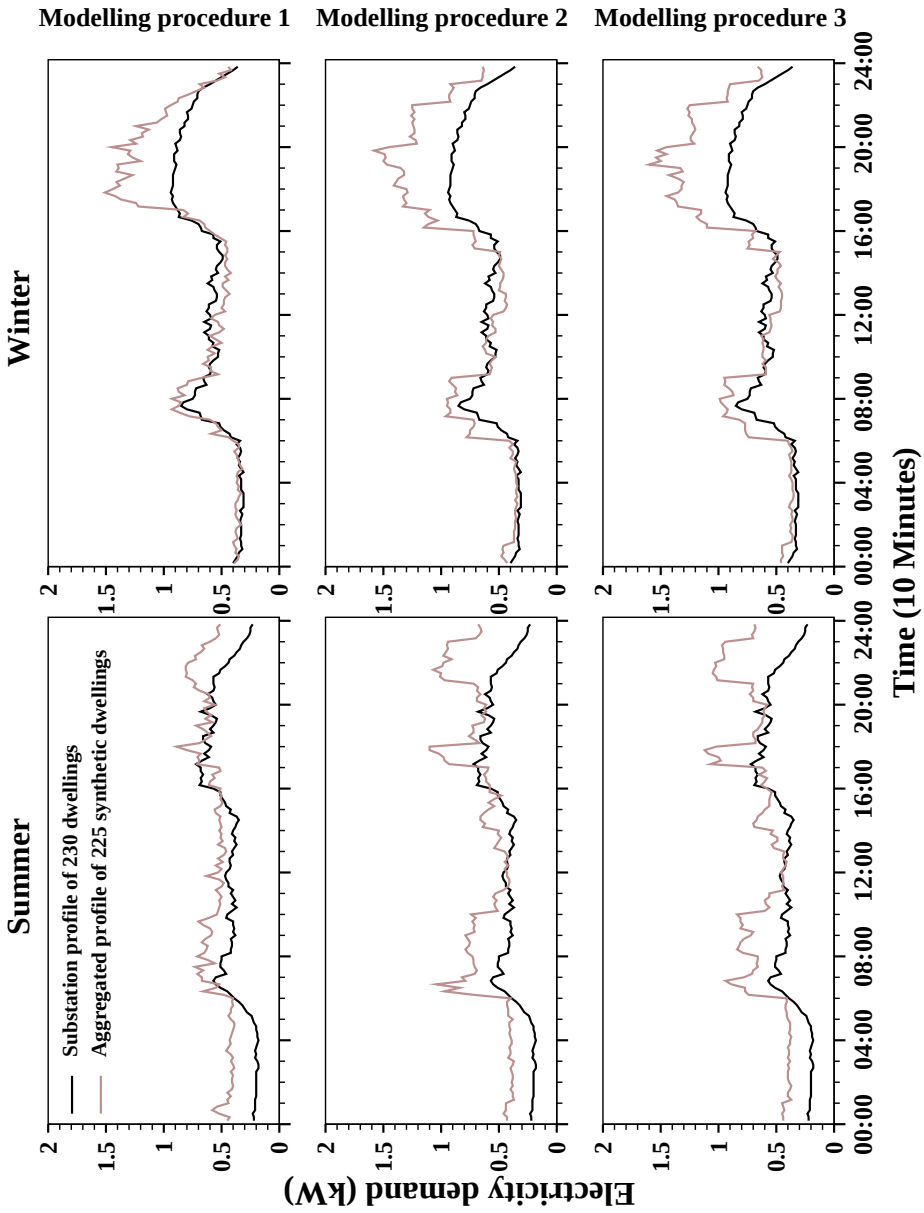

Fig. 7: Substation profile of 230 dwellings [*solid black lines*] at ten-minute resolution compared with aggregated profile of 225 synthetic dwellings [*solid brown line*] for a randomly selected summer day 13^{th} June (**Left Panel**) and for a randomly selected winter day 17^{th} Jan (**Right Panel**). Top row 'modelling procedure 1'; middle row 'modelling procedure 2'; bottom row 'modelling procedure 3'.

References

[1] DEFRA (2012). Climate Change Risk Assessment Summary: Energy.*Department for Environment, Food & Rural Affairs*, Available from http://randd.defra.gov.uk/Default.aspx?Module=More&Location=None&ProjectID=15747 (Accessed August 2016).

[2] Whitesides, G. M., and Crabtree, G. W. (2007). Don't Forget Long-Term Fundamental Research in Energy, *Science*, 315, 796-798.

[3] Jenkins, D.P., Patidar, S. and Simpson, S.A. (2014). Synthesising electrical demand profiles for UK dwellings. *Energy and Buildings*, 76, 605-614.

[4] Patidar, S., Jenkins, D.P. and Simpson, S.A. (23-24 June 2014). Generating Synthetic Enegy Demand Profiles at one minute resolution - A Statistical modelling approach. In *Proceedings of the Building Simulation and Optimization Conference* (Malki-Epsthein, L., Spataru, C., Halburd, L. M. and Mumovic, D. (Ed.), *The Bartlett*, UCL Faculty of the Built Environment, Institute for Environmental Design and Engineering, London, August 2014. ISBN 978-0-9930137-0-6). UCL, London, UK.

[5] Alfares, H. K., and Nazeeruddin, M. (2002). Electric load forecasting: Literature survey and classification of methods. *International Journal of Systems Science*, **33**(1), 23-34.

[6] Walker, C. F., and Pokoski, J. L. (1985). Residential Load Shape Modelling Based on Customer Behaviour. *Power Apparatus and Systems, IEEE Transactions on*, **PAS-104**(7), 1703- 1711.

[7] Capasso, A., Grattieri, Lamedica, W. R. and Prudenzi, A. (1994). A Bottom-up approach to residential load modelling. *IEEE Transactions on Power Systems*, 9, 957-964.

[8] Grandjean, A., Adnot, J. and Binet, G. (2012). A review and an analysis of the residential electric load curve models. *Renewable and Sustainable Energy Reviews*, **16**(9), 6539-6565.

[9] Aguiar, R., and Collares-Pereira, M. (1992). A time-dependent, autoregressive, Gaussian model for generating synthetic hourly radiation. *Solar Energy*, **49**(3), 167-174.

[10] Mora-López, LI. and Sidrach-de-Cardona, M. (1998). Multiplicative ARMA models to generate hourly series of global irradiation. *Solar Energy*, **63**(5), 283–291.

[11] Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (2013). Time series analysis: forecasting and control (4 ed.). *John Wiley and Sons*.

[12] Kwon, H., Lall, U. and Khalil, A.F. (2007). Stochastic simulation model for non-stationary time series using an autoregressive wavelet decomposition: Applications to rainfall and temperature. *Water Resources Research*, 43, W05407.

[13] Schreiber, T., and Schmitz, A. (2000). Surrogate time series. *Physica D*, **142**(3-4), 346–382.

[14] A resampling method for generating synthetic hydrological time series with preservation of cross-correlative structure and higher-order properties. *Water Resources Research*, 48, W12521.

[15] Kitagawa, T., and Nomura, T. (2003). A wavelet-based method to generate artificial wind fluctuation data. *Journal of Wind Engineering and Industrial Aerodynamics*, **91**(7), 943-964.

[16] Hontoria, L., Aguilera, J. and Zufiria, P. (2002) Generation of hourly irradiation synthetic series using the neural network multilayer perceptron. *Solar Energy*, **72**(5), 441–446.

[17] Sanvicente-Sánchez, H., and Solís-Alvarado, Y. (June 30 - July 3, 2008). Generator of Synthetic Rainfall Time Series through Markov Hidden States. In *Computational Science and Its Applications - ICCSA, International Conference Proceedings, Part II*. pp. 959-969, Perugia, Italy.

[18] Silva, E. S., and Hassani, H. (2015). On the use of singular spectrum analysis for

forecasting U.S. trade before, during and after the 2008 recession. *International Economics*, 141, 34-49.

[19] Baum, E. L., and Petrie, T. (1966). Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, **37**(6), 1554-1563.

[20] Rabiner, L. (1989). *A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE*, **77**(2), 257 - 286.

[21] Yamato, J., Jun Ohya and Ishii, K. (15-18 Jun 1992). Recognizing human action in time-sequential images using hidden Markov model. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR '92., 1992 IEEE Computer Society Conference on*, Champaign, IL.

[22] Zhang, Y., Brady, M. and Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *Medical Imaging, IEEE Transactions on.*, **20**(1), pp. 45 - 57, IEEE7.

[23] Eddy, S. R. (1996). Hidden Markov models. *Current Opinion in Structural Biology*, **6**(3), 361-365.

[24] Krogh, A., Brown, M., Mian, S., Sjölander, K. and Haussler, D. (1994). Hidden Markov Models in Computational Biology: Applications to Protein Modeling. *Journal of Molecular Biology*, **235**(5), 1501–1531.

[25] Krogh, A., Larsson, B., Heijne, G. and Sonnhammer, E.L.L. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of Molecular Biology*, **305**(3), 567-580.

[26] Peacock, A. D., and Newborough, M. (2005). Impact of micro CHP systems on domestic sector CO2 emissions. *Applied Thermal Energy*, 25, 2653-2676.

[27] Zucchini, W., and MacDonald, I. L. (2009). *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman and Hall, CRC Press.

[28] Himmelmann, L. (2010). *R package 'HMM'*. Retrieved from http://cran.r-project.org/web/packages/HMM/index.html (Accessed August 2016).

[29] Wild, C. J., and Seber, G. A. (1999). *Chance Encounters: A First Course in Data Analysis and Inference*. (Vol. Chapter 14), Wiley.

[30] Hyndman, R. J. (2015). *R package for ARIMA modelling*. Retrieved from http://cran.r-project.org/web/packages/forecast/index.html (Accessed August 2016).

[31] ARIES (2015). *Adaptation and Resilience In Energy Systems (ARIES)*. project website, http://www.arcc-network.org.uk/project-summaries/aries/#.V7zBlYQrKUk (Accessed August 2016).

Appendix

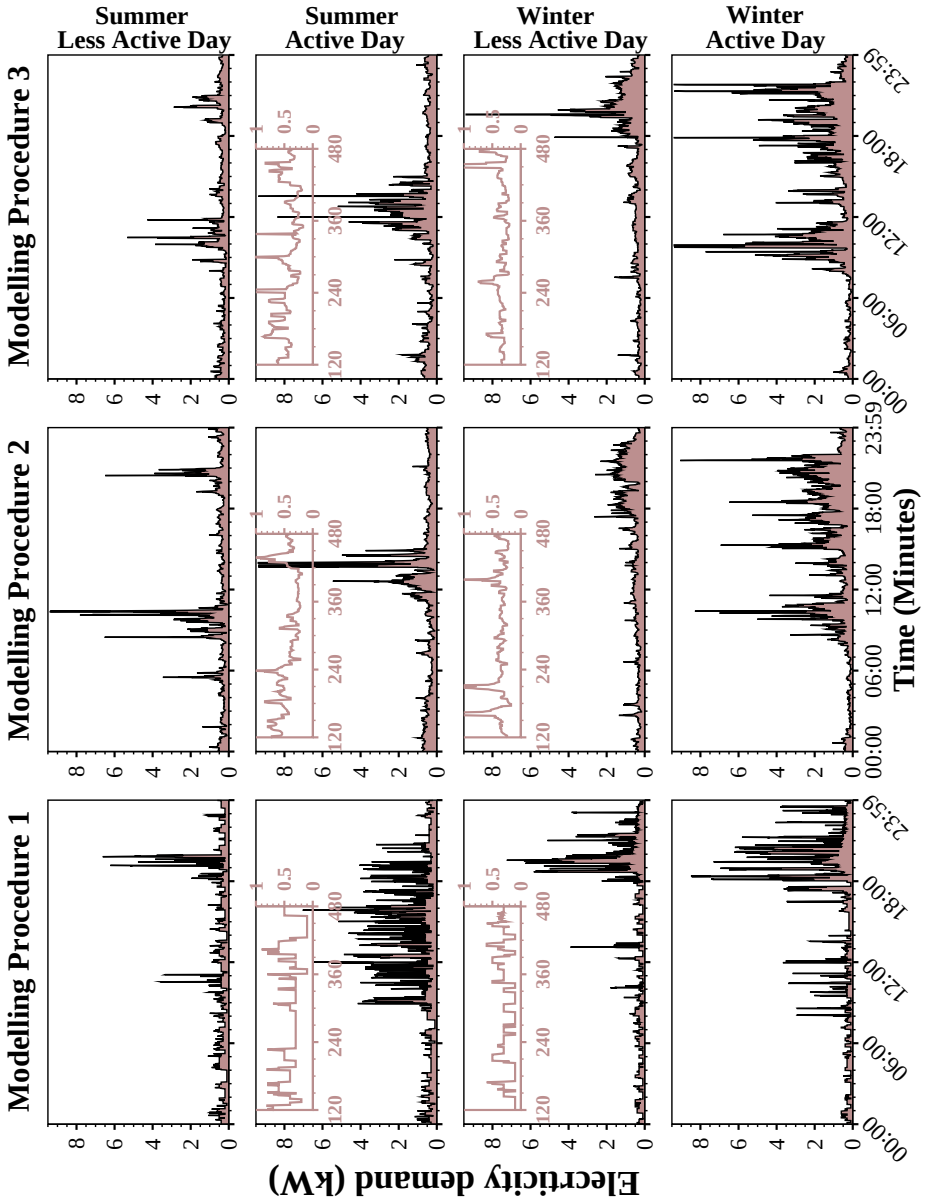

Fig. 8: Snapshot of randomly selected minutely synthetic electricity demand profile for dwelling E generated by three modelling procedures 1, 2 and 3. One active and one less active day have been randomly selected during summer (June-July-August) and winter (November-December-January) period.

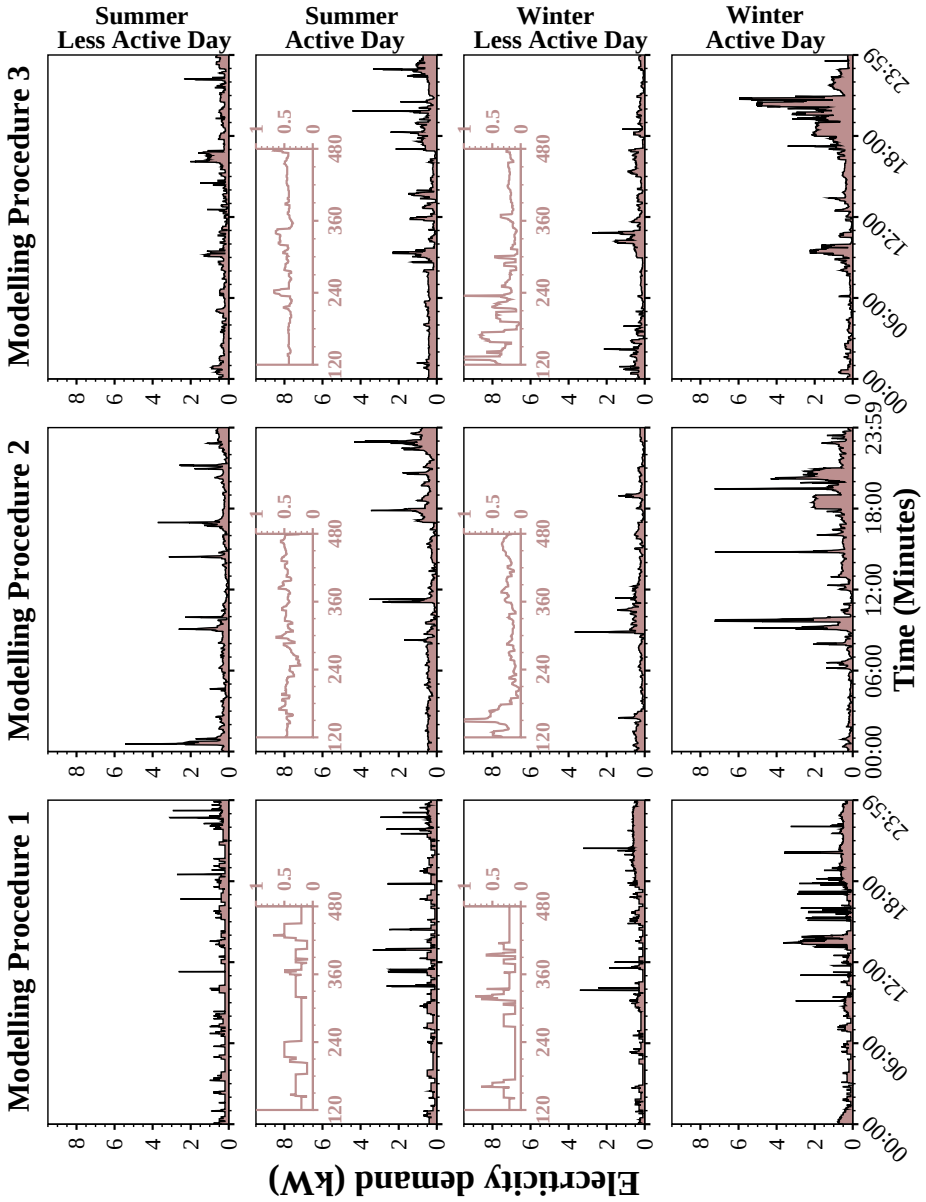

Fig. 9: Snapshot of randomly selected minutely synthetic electricity demand profile for dwelling I generated by three modelling procedures 1, 2 and 3. One active and one less active day have been randomly selected during summer (June-July-August) and winter (November-December-January) period.

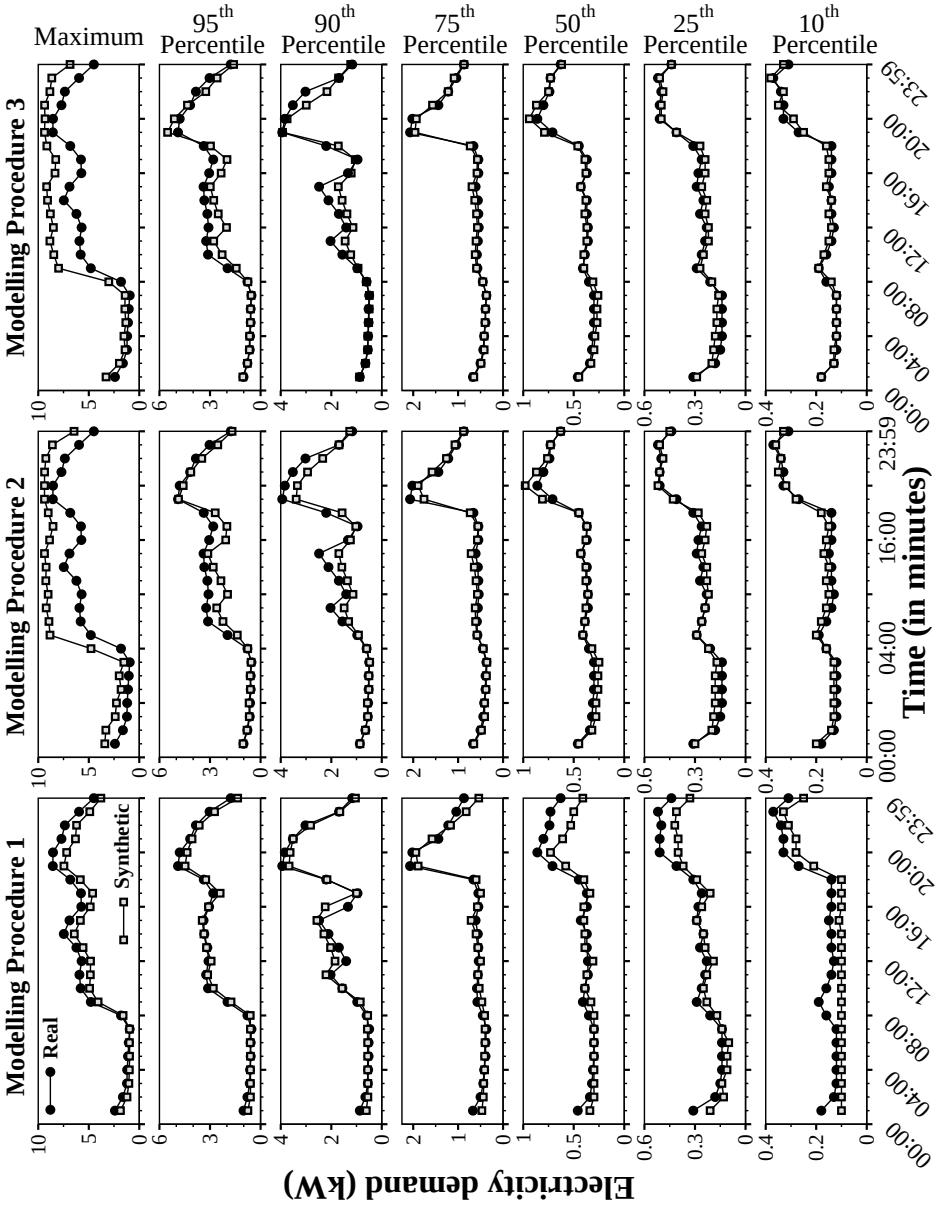

Fig. 10: Maximum and 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th} and 95^{th} percentile at one-hour resolution over the entire year period. Percentiles are compared for real dwelling E with percentiles measured for randomly selected synthetic annual profiles generated by modelling procedure 1, 2 and 3.

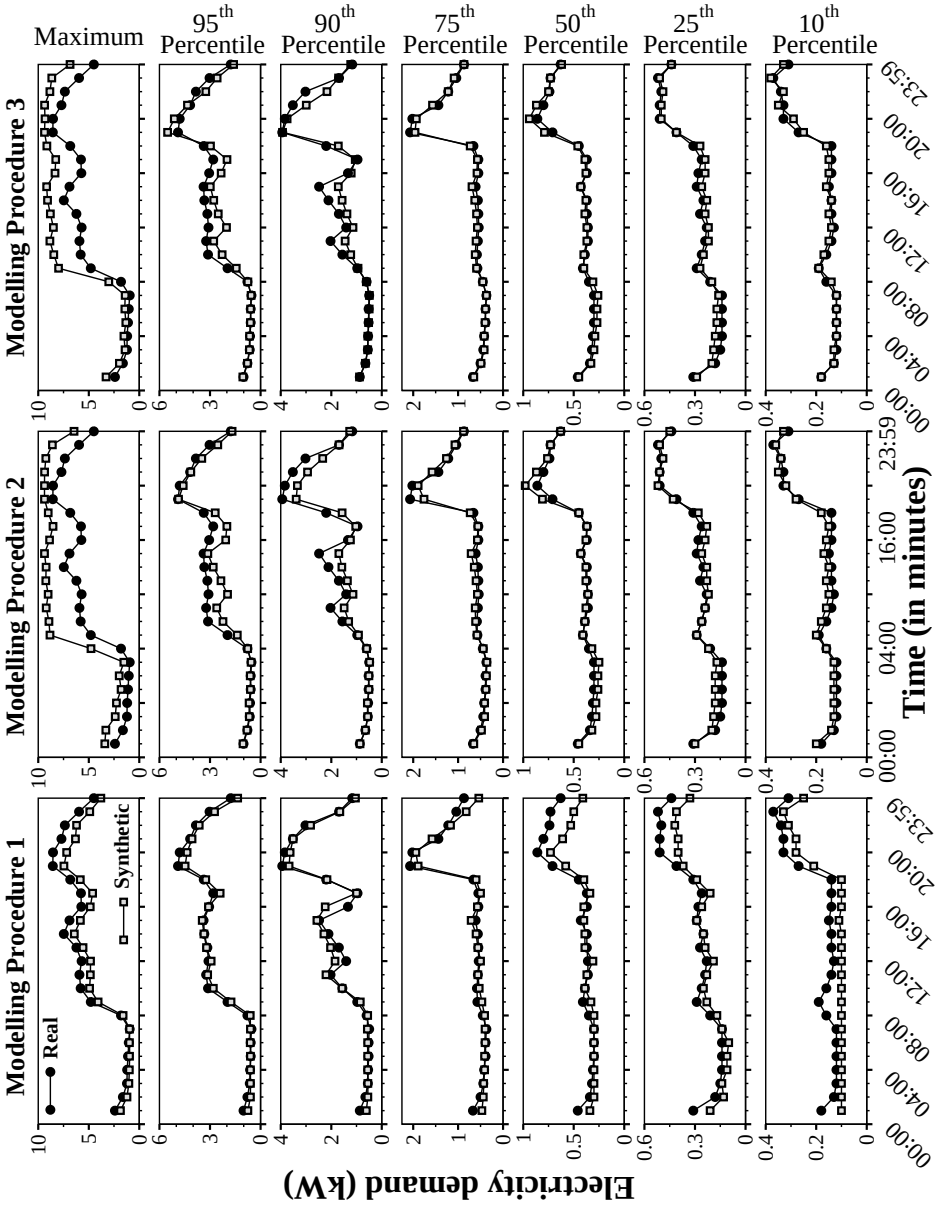

Fig. 11: Maximum and 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th} and 95^{th} percentile at one-hour resolution over the entire year period. Percentiles are compared for real dwelling I with percentiles measured for randomly selected synthetic annual profiles generated by modelling procedure 1, 2 and 3.